Multitask Item Response Models for Response Bias Removal from Affective Ratings

1st Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address

2nd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address

3rd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address

***Abstract*—Response style (RS) is a tendency to choose specific categories regardless of content, e.g. extreme or midpoint categories. It degrades the validity of the analysis of subjective ratings such as correlation and variance-based analyses. However, the computational removal of RS has received little attention from the affective computing community. RS removal techniques have been proposed in areas such as marketing research. However, most of these techniques do not exploit the content-independence of RS; i.e. it should be observed consistently in various tasks, such as affective judgment tasks and standard psychological questionnaires. Therefore, this paper proposes a multitask RS removal method. An individual's responses in multiple tasks are modeled using task-independent RS parameters, and task-dependent parameters, including the item and respondent's characteristic parameters based on item response models (IRM). Through Bayesian modeling, we observed that: i) the proposed model outperformed traditional IRMs in terms of predictive accuracy; ii) our multitask framework estimated RS with higher precision than previous single-task-based RS removal methods; iii) our model replicated Japanese midpoint RS, which has been demonstrated repeatedly in previous cross-cultural studies; and iv) RS-removed predictive ratings showed higher inter-rater agreement than those including RS in valence/arousal judgment tasks.**

I. Introduction

Subjective affect rating still plays an important role in the affective computing community. In fact, the development of effective rating methods is an active research topic [1], [2]. Response styles (RS) are of particular concern when using subjective rating scales. RS is defined as "a systematic tendency to respond to a range of questionnaire items on some basis other than the specific item content (that is, what the items were designed to measure)" [3]. Some of the most common RSs are acquiescent/disacquiescent RSs (ARS/DRS), in which an individual tends to use the upper/lower range of the scale (e.g. yea-saying/nay-saying), and extreme/midpoint RSs (ERS/MRS), in which a person prefers the ends/center of the scale [4]. Traditionally, such RSs were quantized in a simple manner. For example, extreme RSs are frequently measured as the proportion of extreme choices compared with the total number of items [4].

Using such simply calculated measures, researchers have demonstrated how RS degrades the validity of analysis on subjective rating, such as correlation (e.g. correlation between two scales) and variance-based analysis [4]. For example, shared RSs inflate interrater agreement [5], and deflates it if RSs differ between raters [6]. It is serious in cross-cultural studies because of the cultural differences in RS. For example, American college students tended to have more extreme RS than Japanese students [7], [8]. Traditional methods provide measurement about RS, but they cannot eliminate RS from ratings.

Recently, several ways to remove RS have been proposed. The first way is to obtain multiple ratings for the same target and to aggregate them, assuming random independent noise across the ratings. The target may be ratings from multiple people to a visual/auditory stimulus, or ratings from a single person to a set of items about a psychological construct (e.g. psychological questionnaire). This type of studies covers both simple aggregation methods, including averaging and majority voting, and more advanced truth discovery methods [9]–[11]. This technique is useful when the research target is perceived emotion, i.e. emotion perceived by general population. However, if the target is individual-level perception, such as felt emotion (what the target person is actually feeling) or how another specific individual perceives it, then such techniques are difficult to apply because of their high cost and/or reproducibility. The second method is to use anchors in rating to correct individual differences in the criteria for selecting a category. Example methods are ordinal rating [12], and Anchoring Vignette method [6]. However, such anchors should be carefully designed specific to the target task; the designing per se remains a research topic.

The third way to eliminate RS is to build a rating process model with RS as a latent variable and remove the effect of RS. Several RS removal techniques have been proposed mainly in the marketing research area [6], [13]. Most of them use a single task. For example, when emotion judgment is the target task, RS is estimated only by using ratings on the task. The main weakness of this approach is that it is difficult to distinguish RS from task-dependent response tendencies; the two types are called dispositional and situational in [4]. For example, depending on the item/stimulus and/or prepared categories to choose, people may tend to select extreme responses in some tasks (e.g. because of many exaggerated facial expressions in

Identify applicable funding agency here. If none, delete this.

an affect rating task), while middle category in other tasks (e.g. as a consequence of the ambiguity of items). In such cases, the single task framework yields different results of RS, which unfortunately violates the definition of RS.

Focusing on the task independence of RS, we propose to use various tasks together to extract RS shared across tasks. In fact, in many affective computing studies, Likert-type psychological questionnaires are additionally used to examine the relationship between the results of main task and the summary statistics (primarily total score) of the additional tasks. Modern test theories, including item response theory (IRT), make it possible to estimate both individual's characteristics and each item's characteristics jointly from answers to such questionnaires. Therefore, we propose a multitask item response model for RS removal.

To the best of our knowledge, this paper has two major contributions. First, this is the first attempt in the affective computing community to computationally remove RS in affective ratings. Second, it is the first IRM-based multitask framework for estimating/removing RS. This paper demonstrates how our multitask framework works in one of the most fundamental tasks in the affective computing area, namely valence and arousal judgment tasks. The proposed framework can potentially open new horizons for future affective computing studies.

II. Method

Our model is an extension of item response models (IRMs) that contain response style (RS) parameters for polytomous ratings. This section first introduces basic IRMs which have no RS terms, and more advanced IRM with RS. Next, our model is explained.

A. Single task models

1) Basic item response models: IRM is a family of multivariate generalized linear mixed models (MGLMM) [14]. IRM consists of three parts: 1) distribution of data, 2) link function, which determines what transformation of the mean of the distribution to be modeled linearly, and 3) predictors.

In common IRMs, multivariate Bernoulli distribution, namely a multinomial distribution with total count equal to one, is used with an adjacent-categories logit. Such IRMs are expressed as

$$\log(\frac{P(y_{ij}=s|\boldsymbol{X}_\Theta)}{P(y_{ij}=s-1|\boldsymbol{X}_\Theta)})=\boldsymbol{X}_\Theta \tag{1}$$

where y_{ij} denotes the response of person j to item i, and $\boldsymbol{X}_\Theta$ is a linear predictor that consists of a set of parameters Θ including person (respondent) parameter and item (stimulus for affective judgment or item in psychological questionnaire) parameter.

One of the most fundamental item response model is Partial Credit Model (PCM) [15], in which $\boldsymbol{X}_\Theta$ is defined to be $\theta_j-\beta_{is}$. θ_j is the trait of person j (such as ability in the test theory domain), while β_{is} is the characteristics of item i for category s (e.g. the difficulty of the item to obtain score s or selection threshold/criterion for rating s). Generalized PCM (GPCM) [16] is an extended version of PCM, where the effect of person ability is assumed to be different across items; namely $\boldsymbol{X}_\Theta=\alpha_i\theta_j-\beta_{is}$, where $\alpha(>0)$ is called slope parameter (or discrimination parameter), because it determines the slope of characteristic curve (a cumulative distribution) that represents the relationship between the probability of obtaining the category and the individual's ability. Both models, like many other IRMs, ignore any temporal structure, and assume that the model does not change over time.

One of the key properties of PCM is that it inherits from the specific objectivity property of the original Rasch model; that is, the comparison of items does not depend on person parameters, and the comparison of persons does not depend on item parameters [13]. When two items i and i' are compared for the same person j, the difference of their logits has no person term:

$$\begin{aligned}\log(\frac{P(y_{ij}=s|\boldsymbol{X}_\Theta)}{P(y_{ij}=s-1|\boldsymbol{X}_\Theta)})-\log(\frac{P(y_{i'j}=s|\boldsymbol{X}_\Theta)}{P(y_{i'j}=s-1|\boldsymbol{X}_\Theta)})\\ =(\theta_j-\beta_{is})-(\theta_j-\beta_{i's})=\beta_{is}-\beta_{i's}.\end{aligned} \tag{2}$$

This property also holds for the difference between two persons for the same item. On the other hand, GPCM does not preserve the property due to the interaction term $\alpha_i\theta_j$.

2) Response style models for single task: Recently, several researchers have proposed to incorporate RS into traditional IRMs. One example was proposed by Tutz et al. [13] who incorporated RS term $\tilde{\gamma}$ into threshold β_{is} as $\tilde{\beta}_{is}=\beta_{is}-\tilde{\gamma}_{js}$ (which this paper calls PCM_RSt):

$$\boldsymbol{X}_\Theta=\theta_j-(\beta_{is}-\tilde{\gamma}_{js}), \tag{3}$$

where $\tilde{\gamma}_{js}=(m-s+1)\gamma_j$, m is the midpoint category (e.g. $m=2$ when $s\in\{0,1,2,3,4\}$ and $m=2.5$ when $s\in\{0,1,2,3,4,5\}$). Positive γ represents midpoint RS, while negative γ means extreme RS. If γ is positive, the intervals of β between categories expands around the middle category m, which means that the probability of category m increases (i.e. midpoint RS). If γ is negative, it has the opposite effect, namely the intervals go toward the middle category, and consequently the probability of extreme categories increases (i.e. extreme RS).

On the other hand, Jonas and Markon [6] incorporated extreme/midpoint RS and positive/negative bias into the GPCM (which this paper calls GPCM_RSj) as

$$\boldsymbol{X}_\Theta=\alpha_i\theta_j-\gamma_j(\beta_{is}-\gamma'_j). \tag{4}$$

γ represents extreme/midpoint RS, as in Tutz et al.'s model (although in Jonas & Markon's model, $\gamma>0$ and smaller/larger value means extreme/midpoint RS), while γ' represents a bias toward positive/negative category representing acquiescent/disacquiescent RS. Tutz et al.'s model satisfies the specific objectivity property because it has no interaction between person and item parameters. On the other hand, Jonas & Markon model does not satisfy it due to the interaction term.

TABLE I
LIST OF FAMILY OF ITEM RESPONSE MODELS

Model	Predictors $\boldsymbol{X}_\Theta$
Models w/o response style	
Baseline models	
PCM [15]	$\theta_{jk} - \beta_{iks}$
GPCM [16]	$\alpha_{ik}\theta_{jk} - \beta_{iks}$
Models w/ response style	
Baseline models	
PCM_RSt [13]	$\theta_{jk} - (\beta_{iks} + \tilde{\gamma}_{jks})$
GPCM_RSj [6]	$\alpha_{ik}\theta_{jk} - \gamma_{jk}(\beta_{iks} + \gamma'_{jk})$
Proposed models	
mtPCM_RSt	$\theta_{jk} - (\beta_{iks} + \tilde{\gamma}_{js})$
mtGPCM_RSt	$\alpha_{ik}\theta_{jk} - (\beta_{iks} + \tilde{\gamma}_{js})$
mtPCM_RSj	$\theta_{jk} - \gamma_j(\beta_{iks} + \gamma'_j)$
mtGPCM_RSj	$\alpha_{ik}\theta_{jk} - \gamma_j(\beta_{iks} + \gamma'_j)$

All model use adjacent-categories logit as link function. β is an item parameter subscripted with item index i (and category index s for some cases). θ is a person parameter subscripted with person index j (and category index s for some cases). α is a scale parameter subscripted with item index i. k is a task index. Note that we include task index k also to baselines for comparison.

B. Proposed multitask models

We extend Tutz et al.'s [13] and Jonas & Markon's [6] models to a multitask framework. We incorporate a set of tasks all together using task-independent parameters that describe RS. Our multitask version of Tutz et al.'s model (mtPCM_RSt) is defined as:

$$\boldsymbol{X}_\Theta = \theta_{jk} - (\beta_{iks} + \tilde{\gamma}_{js}), \tag{5}$$

for where k is a task index. Note that $\tilde{\gamma}$ excludes subscript k. Our extension of Jonas & Markon's model (mtGPCM_RSj) is:

$$\boldsymbol{X}_\Theta = \alpha_{ik}\theta_{jk} - \gamma_j(\beta_{iks} + \gamma'_j). \tag{6}$$

We also built GPCM version of mtPCM_RSt and PCM version of mtGPCM_RSj by replacing θ_{jk} and $\alpha_{ik}\theta_{jk}$ (called mtGPCM_RSt and mtPCM_RSj, respectively). The models based on Tutz et al.'s model (namely mtPCM_RSt and mtGPCM_RSt) satisfy the specific objectivity property, while those based on Jonas & Markon's model (mtPCM_RSj and mtGPCM_RSj) do not. Table I compares all four proposed models with the baseline models.

After estimating model parameter Θ, we can predict the ratings that are likely to be drawn from the model. Predictive rating $\hat{y}$ is estimated as:

$$\hat{y}_{ijk} \sim categorical(\boldsymbol{\pi}) \tag{7}$$

$$\pi_s = P(y = s|\boldsymbol{X}_\Theta) \tag{8}$$

In addition, the RS-removed ratings are estimated by excluding the RS term $\tilde{\gamma}$ from the predictor in Eq. 8. This can be expressed as:

$$\pi_s = P(y = s|\boldsymbol{X}_{\Theta'}) \tag{9}$$

where $\boldsymbol{X}_{\Theta'} = \theta_{jk} - \beta_{iks}$ for the PCM families and $\boldsymbol{X}_{\Theta'} = \alpha_{ik}\theta_{jk} - \beta_{iks}$ for the GPCM families.

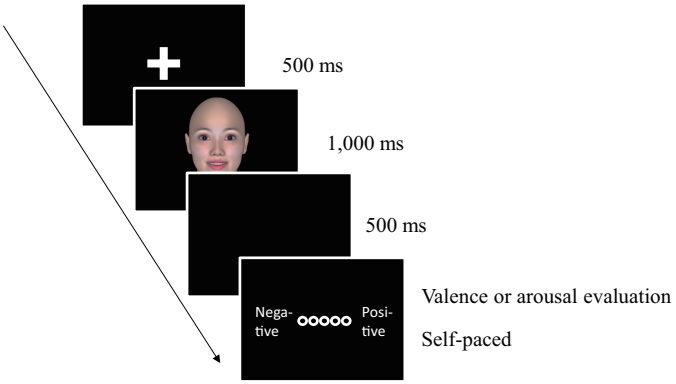


Fig. 1. Main task: valence-arousal judgment task. First, a fixation cross was displayed at the center for 500 msec. Next, a target face was shown for 1,000. Then, it disappeared for 500 msec. Finally, valence or arousal scale was displayed until participant selected one of the answers.

III. EXPERIMENTAL DATA

In order to evaluate the proposed framework, we performed valence and arousal judgment tasks using computer-generated static emotional faces as main tasks. We also used psychological questionnaires as subtasks.

A. Observers

Fifty Japanese university students (25F) participated in the experiment. This homogeneity facilitates the verification of whether the estimated RS parameters really match the Japanese midpoint RS, which is repeatedly reported in previous studies [7], [8].

B. Main tasks: affective rating

The participants were asked to rate the valence and arousal level of artificial faces. This was a blocked design: one block for valence judgment and the other block for arousal judgment. Each was a forced choice on a 5-point scale: the extremes were labeled "Positive" and "Negative" in the valence block, while "High" and "Low" in the arousal block. Figure 1 shows the timeline of each trial. Each block consisted of 150 trials. In 120 of 150 trials, totally 120 original faces were displayed. The remaining 30 trials were repetition of 30 trials that were randomly selected out of the 120 trials. This was aimed at calculating the test-retest reliability, i.e. the frequency with which participants gave the same rating to exactly the same face in different trials. The block order was counter-balanced, and the stimulus order and the 30 repeated faces in each block were randomized across the participants. All the labeling was done in isolation, and all the observers successfully completed both tasks.

Various mixed facial expressions were included to see inter-individual differences in perceptions among respondents. The 120 stimulus faces were created using FaceGen modeler. The faces consisted of 29 facial expressions (neutral and 28 non-neutral expressions) from 8 different artificial identities. Specifically, 15 expressions (neutral and 14 non-neutral expressions) from 4 virtual identities (called Face Set 1), and

TABLE II
SUMMARY OF USED TASKS

Task	#items/trials	#points
Main tasks		
1. Valence judgment	150	5
2. Arousal judgment	150	5
Sub-tasks		
3. EQ [17]	60	4
4. SQ [17]	60	4
5. AQ [18]	50	4
6. IRI [19]	28	3
7. ESCQ [20]	28	4
8. B5 [21]	60	5
9. TEG [22]	53	7
Sum	639	

neutral and the remaining 14 non-neutral expressions (totally 15 expressions) from the other 4 identities (called Face Set 2) were extracted. Expressions were manipulated by changing the modeler's expression-specific parameters (angry, disgust, fear, sad, surprise, and closed- and open-mouth smiles; totally seven categories). Of the non-neutral expressions in Face Set 1, four were pure angry, fear, surprise and open-mouth smile, and the remaining 10 were combinations of the seven categories. Three of the non-neutral expressions of Face Set 2 were pure disgust, sad and closed-mouth smile, and the remaining 11 were other combinations of the seven categories. The eight identities were from Caucasians, Africans, Indians and Asians: each of which consists of both masculine and feminine faces. This procedure yielded 120 (=15×4+15×4) faces.

C. Subtasks: psychological questionnaires

The participants were also asked to answer seven psychological questionnaires after the main tasks: Empathizing Quotient (EQ) [17], Systemizing Quotient (SQ) [17], Autism-Spectrum Quotient (AQ) [18], Interpersonal Reactivity Index (IRI) [19], Emotional Skills and Competence Questionnaire (ESCQ) [20], Neo-FFI or Big Five (B5) [21], and Tokyo University Egogram (TEG) [22]. EQ, AQ, IRI and ESCQ are commonly used to measure empathy-related traits, while B5 and TEG are for more general personality traits. They are not completely independent of each other, nor are they fully independent of valence/arousal decision tasks. However, the entire questionnaire set reasonably covers various types of traits and the number of points (ranging from 3-point scale to 7-point scale, and including both even and odd points).

Table II summarizes the number of items and the number of points of the questionnaires. The total number of ratings was 579 items $\times$ 50 respondents $= 31,950$. There was no missing data. However, our models accept missing data in the current form, thanks to Bayesian generative modeling as described in IV-A.

IV. EVALUATION SETTINGS

A. Bayesian parameter estimation

All the models were implemented using the free and open-source software Stan and its interface to the R (Stan Development Team, 2015a, b), and the edstan (v1.0.6; Furr, 2017) package. The model parameters were estimated using Stan's No-U-Turn Sampler (NUTS). As weak priors, zero-mean normal distributions were used for β, θ and γ (for Tutz et al's families), unit-mean lognormal distributions were used for α and γ (for Jonas & Markon families). Four MCMC chains were run from random start values. The chain convergence was assessed by the $\hat{\mathrm{R}}$ statistic ($\hat{\mathrm{R}} < 1.1$). The first 2,400 iterations were discarded as warm-up, and then 2,400 iterations were obtained and stored from each chain, yielding 9,600 iterations that served to empirically approximate the posterior distribution.

Predictive RS-inclusive ratings and RS-removed ratings ($\hat{y}$) were obtained as follows. The ratings for all respondents and items (31,950 samples) were generated (simulated) according to Eq. 7 and Eq. 8 for RS-inclusive ratings, and Eq. 9 for RS-removed ratings. This procedure was repeated 9,600 times. This yielded the posterior distribution consisting of 9,600 random samples for each $\hat{y}$ for both types of predictive ratings. A point estimate was further determined for each $\hat{y}$ by majority voting by the 9,600 samples. The following analysis used the point estimates unless otherwise specified.

B. Performance measure

As evaluation criteria for model comparison, the approximate widely applicable information criterion (WAIC) [23] and Pareto smoothed importance sampling leave-one-out cross-validation (PSIS-LOO) [24] (an approximated LOO) were calculated using the loo package (v.2.0.0; https://mc-stan.org/loo/). Both measures penalize model complexity, and a smaller value indicates a better model. We also report the following four measures for the main valence and arousal tasks to indicate how well each model explains the observed ratings: accuracy (percent agreement, κ), Pearson's correlation coefficient (r), mean absolute error (MAE), and intra-class correlation coefficient (ICC), following [25], which recommend the use of multiple measures jointly.

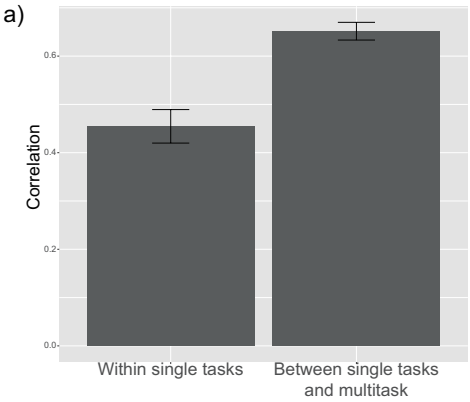
V. RESULTS

This section reports various validation results, including model comparison, and prediction performance evaluation. All the results support the validity of our proposed framework.

A. Rating results

1) Basic statistics: The proportion of the rating categories (the marginal distribution of ratings) was (09, .32, .35, .20, .04) (from negative to positive) for valence, and (.09, .25, .31, .28, .07) (from low to high) for arousal.

The test-retest reliability (calculated in a manner similar to accuracy) κ was .525 for valence and .475 for arousal. This is a percent agreement, meaning that the participants gave the same rating between the test and retest pairs at a rate of κ. Fleiss' generalized κ, κ_F, a chance-corrected agreement, was .345 and .300 for valence and arousal rating. Pearson's r was .556 for valence and .550 for arousal. This value is comparable to that reported in the literature, e.g. [26]. The ICC(2,1) was .48 for valence and .35 for arousal. Both are considered to be



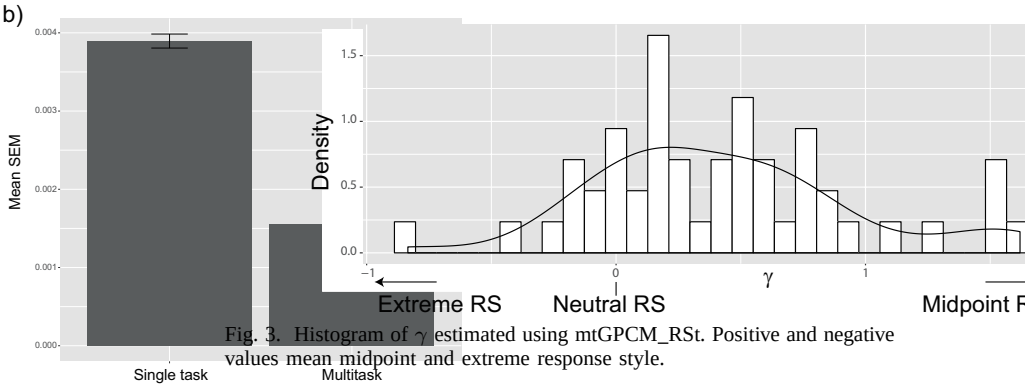


Fig. 2. (a)Correlation of estimated γ within single tasks (using GPCM_RSt) and those between single tasks (using GPCM_RSt) and multitask (using mtGPCM_RSt). (b) Mean of SEM of γ's posterior distribution: GPCM_RSt vs mtGPCM_RSt. Error bar indicates SEM.

Fig. 3. Histogram of γ estimated using mtGPCM_RSt. Positive and negative values mean midpoint and extreme response style.

between poor and fair [27], [28]. These well demonstrate how differently people rate affective faces.

However, it is uncertain whether it is caused by the individual difference of perception or by the RS. Therefore, we investigate the impact of RS on these reliability measures in V-D.

B. Model comparison

Table III summarizes model performance. PCM_RSj [6] and mtGPCM_RSj did not converge on learning ($\hat{R} > 4$), thus were excluded from the following analyses. In terms of all the criteria, our best model (mtGPCM_RSt) outperformed the baselines (PCM and GPCM)[1]. As the base model, GPCM was preferred to PCM; mtGPCM_RSt was slightly better than mtPCM_RSt (78,008 vs 78,069 for WAIC, and 78,212 vs 78,274 for LOO).

There found severe bugs in the source codes in multitask models. 1) The numbe of betas is much larger than necessary number. 2) Theta should be task-dependent! Currently, it is assumed to be task independent. The vertical axis of Figure 2 (b) is wrong. It should be Mean SEM etc. Cohen's d cannot be used for one-sample t-test.

[1]The accuracy of mtGPCM_RSt was higher than the test-retest reliability. It may sound strange, but it is possible. The *upper bound* of the prediction accuracy is estimated to be .83 for valence and .81 for arousal. The upper bounds were obtained as follows. Our data contains two types of data and they should be considered separately. Of the 120 images (the 150 trials), 30 (60) were shown twice, and the remaining 90 (90) were used only once. For the 60% (=90/150) samples, the perfect accuracy is possible if a very complex model is used (although probably overfitting). This is because the training and test sets were identical. For the 40% (=60/150) samples, κ_F percent of samples, where the test and retest ratings are identical (not by chance), the perfect accuracy is also possible. The remaining $1-\kappa_F$ percent of samples were however rated differently between a pair of trials, and thus the perfect accuracy is not possible. This is because the proposed models (as well as the baselines) give the same predictive rating for each pair of trials. If random sampling of rating from the marginal distribution is assumed for the samples, the maximum chance level (p_{max}) is .35 and .31 for valence and arousal tasks, respectively. Therefore, the estimated upper bound for the 40% data is $\kappa_F \times 1 + (1-\kappa_F) \times p_{max} = .574$ for valence and .517 for arousal. Taken together, the overall upper bound is expected to be $.574 \times 40\% + 1 \times 60\% = 0.83$ for valence, and $.517 \times 40\% + 1 \times 60\% = 0.81$ for arousal. The observed accuracies are within the range.

C. Single-task vs multitask

In Table III, PCM_RSt [13] outperformed our mt(G)PCM_RSt. This means that if the objective is to describe the observed ratings as accurately as possible, PCM_RSt should be selected. However, as mentioned in I, single task framework confuses task-dependent response tendencies with RS. To further illustrate the need for the multitask framework quantitatively, Fig. 2 (a) shows the pairwise correlation of estimated γ (a 50-d vector) within the 9 single tasks using GPCM_RSt (yielding a within-single-task correlation for each of ${}_9C_2 = 36$ pairs of tasks). It also includes the correlation between the γ values and those obtained using our mtGPCM_RSt. The estimated γ in single task was closer to the estimate in the multitask than that in a different single task. It reasonably demonstrates the task-independence of RS.

In addition, Fig. 2 (b) shows another benefit of using multiple tasks; the multitask framework gave more precise estimate. The posterior distribution of γ was narrower in the multitask scenario (mtGPCM_RSt) than in the single scenarios (GPCM_RSt). This is an important property because γ parameters are interconnected with the other parameters and thus precise estimate of γ is expected to lead to precise estimates of the remaining parameters.

D. Estimated parameters and response style removal

Figure 3 shows a histogram of estimated γ across 50 participants using GPCM_RSt. Here we use GPCM_RSt, not the best model, because γ of PCM_RSj has no clear threshold between extreme and midpoint RSs. The mean value was positive (M = 0.42 ($\pm$0.07 SEM), $p < .001$, d = .81), indicating that the participants had midpoint RS overall. The midpoint RS of Japanese people is in line with previous studies [7], [8]. Furthermore, the estimated γ values were reasonably correlated with the traditional measure of extreme RS, i.e. the proportion of extreme choices out of the whole items [4] (Spearman's $\rho = -.91, p < .001$). These results validate our method. Moreover, γ of GPCM_RSt and γ of PCM_RSj showed strong correlation; $\rho = .78, p < .001$.

An ICC(2,1) of the estimated posterior ratings obtained by Eqs. 7 and 8 for the whole 9,600 samples (not their point estimates), namely a recovered ICC, was M = .49 (95% CI [.47, .51]) for valence and M = .36 (95% CI [.34, .38] for

TABLE III
PREDICTIVE PERFORMANCE OF THE PROPOSED MODELS AND BASELINES FOR THE WHOLE NINE TASKS

Model	WAIC ↓		LOO ↓		Valence task				Arousal task			
	Mean	SEM	Mean	SEM	$\kappa \uparrow$	$r \uparrow$	MAE ↓	ICC ↑	$\kappa \uparrow$	$r \uparrow$	MAE ↓	ICC ↑
PCM [15]	78,771	272	78,957	277	.583	.643	.484	.637	.452	.528	.692	.523
GPCM [16]	77,016	277	77,273	283	.585	.644	.481	.637	.453	.537	.688	.533
PCM_RSt [13]	73,683	284	73,870	289	.630	.664	.443	.661	.499	.567	.646	.566
GPCM_RSt	72,438	287	72,698	292	.630	.664	.443	.661	.499	.567	.646	.566
mtPCM_RSt	76,589	279	76,773	284	**.601**	.650	.468	.645	.470	.546	.674	.543
mtGPCM_RSt	75,126	283	75,361	288	**.601**	.650	.468	.645	.470	.546	.674	.543
mtPCM_RSj	**72,117**	245	**72,345**	248	**.601**	**.652**	**.467**	**.647**	**.472**	**.551**	**.670**	**.548**
mtGPCM_RSj	74,881	240	74,949	241	.582	.646	.480	.638	.450	.531	.686	.525

"↑" and "↓" denote higher and lower performance. Note that although achieving the best performance in terms of the predictive performance of *ratings*, PCM_RSt [13], a single task framework, confuses task-dependent response tendencies with RS, as mentioned in I.

arousal. The observed ICCs (.48 and .35, respectively) were successfully replicated.

The RS-removed ratings were estimated following Eqs. 7 and 9. This slightly but statistically significantly increased recovered ICC; .51 (95% CI [.49, .54]) for valence and .41 (95% CI [.38, .44]) for arousal. This suggests that in our participant set, the observed ICCs were deflated because many participants had midpoint RS while some had extreme RS. This supports the need for RS correction.

VI. DISCUSSION

We have provided a variety of evidence in support of our multitask framework. However, several issues still remain.

First, our multitask framework successfully found Japanese midpoint RS. However, this was an indirect evaluation, and a more direct evaluation is needed. One way is to use an anchoring vignette technique, such as [6], in which respondents are also asked to judge imaginary character(s) as *anchor* that are assumed to cause the same judgment across people, in order to normalize each respondent's judgment based on their judgment on the anchor.

Second, our model is probably not the *best model* to eliminate RS in a multitask fashion. First, although we use the same base model (PCM or GPCM) for all tasks, we can use different models for different tasks in our framework. It is reasonable to use a simple model (e.g. PCM) for psychological questionnaires, since they are basically designed to measure a single construct. However, it would be interesting to find the best, or at least better, model for affective judgment tasks.

Thirdly, this study employed a discrete annotation procedure for both time and emotion space. To apply our work to continuous annotations, as with the recent trend in the affective community, e.g. [1], [2], our model must be extended. It is also interesting to investigate whether the rating process is time invariant or not, as mentioned in [12].

Finally, this paper focused on decoder or receiver in emotional communication, i.e. affective judgment to other people. It is also interesting to target coder's or sender's judgment, i.e. self-report of emotional states. This is an important step because self-report is available only from the individual. Therefore, the impact of RS on their ratings is expected to be stronger than that of decoder. It would be interesting to incorporate our framework with physiological signals, as used in felt-emotion studies [29]–[31].

VII. CONCLUSION

This paper proposed a multitask RS removal framework, where individual's responses in multiple tasks are modeled using task-independent RS terms, and task-dependent terms, including item and respondent's characteristic parameters based on item response model (IRM). Through a Bayesian modeling, we observed that i) the proposed model outperformed traditional IRMs in terms of predictive accuracy; ii) our multitask framework estimated RS with higher precision than previous single-task-based RS removal methods; iii) our model replicated Japanese midpoint RS, which has been repeatedly shown in previous cross-cultural studies; and iv) RS-removed predictive ratings showed higher inter-rater agreement than those including RS in valence/arousal judgment task. The proposed RS removal technique has a potential to reveal stronger/new results that previous methods could not find in the affective computing community. Validating the potential is one of the next steps.

REFERENCES

[1] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schröder, "FEELTRACE: An instrument for recording perceived emotion in real time," in *ISCA Workshop on Speech and Emotion*, 2000, pp. 19–24.

[2] G. N. Yannakakis and H. P. Martínez, "Grounding truth via ordinal annotation," in *Proc. ACII*, 2015, pp. 574–580.

[3] D. L. Paulhus, *Measures of personality and social psychological attitudes*. San Diego, CA, US: Academic Press, 1991, ch. Measurement and control of response bias, pp. 17–59.

[4] H. Baumgartner and J.-B. E. Steenkamp, "Response styles in marketing research: A cross-national investigation," *Journal of Marketing Research*, vol. 38, no. 2, pp. 143–156, 2001.

[5] S. Dolnicar and B. Grün, "Response style contamination of student evaluation data," *Journal of Marketing Education*, vol. 31, no. 2, pp. 160–172, 2009.

[6] K. G. Jonas and K. E. Markon, "Modeling response style using vignettes and person-specific item response theory," *Applied Psychological Measurement*, vol. 43, no. 1, pp. 3–17, 2019.

[7] M. Zax and S. Takahashi, "Cultural influences on response style: Comparisons of japanese and american college students," *The Journal of Social Psychology*, vol. 71, no. 1, pp. 3–10, 1967.

[8] C. Chen, S.-Y. Lee, and H. W. Stevenson, "Response style and cross-cultural comparisons of rating scales among east asian and north american students," *Psychological Science*, vol. 6, no. 3, pp. 170–175, 1995.

[9] A. P. Dawid and A. M. Skene, “Maximum likelihood estimation of observer error-rates using the EM algorithm,” *Appl. Statist.*, vol. 28, no. 1, pp. 20–28, 1979.

[10] P. Welinder, S. Branson, S. Belongie, and P. Perona, “The multidimensional wisdom of crowds,” in *Neural Information Processing Systems Conference (NIPS)*, 2010, pp. 2424–2432.

[11] A. Rui, O. Martinez, X. Binefa, and F. Sukno, “Fusion of valence and arousal annotations through dynamic subjective ordinal modelling,” in *Proc. IEEE FG*, 2017 2017.

[12] G. N. Yannakakis, R. Cowie, and C. Busso, “The ordinal nature of emotions,” in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2017, pp. 248–255.

[13] G. Tutz, G. Schauberger, and M. Berger, “Response styles in the partial credit model,” *Applied Psychological Measurement*, vol. 42, no. 6, pp. 407–427, 2018. [Online]. Available: https://doi.org/10.1177/0146621617748322

[14] F. Tuerlinckx and W.-C. Wang, *Explanatory Item Response Models*. New York, NY: Springer, 2004, ch. Models for polytomous data, pp. 75–109.

[15] G. N. Masters, “A rasch model for partial credit scoring,” *Psychometrika*, vol. 47, no. 2, pp. 149–174, Jun 1982.

[16] E. Muraki, “A generalized partial credit model: Application of an em algorithm,” *ETS Research Report Series*, vol. 1992, no. 1, pp. i–30, 1992.

[17] S. Baron-Cohen, “Autism: The empathizing-systemizing (e-s) theory,” *Ann. N. Y. Acad. Sci.*, vol. 1156, pp. 68–80, 2009.

[18] S. Baron-Cohen, S. Wheelwright, R. Skinner, J. Martin, and E. Clubley, “The autism-spectrum quotient (aq): Evidence from asperger syndrome/high-functioning autism, malesand females, scientists and mathematicians,” *Journal of Autism and Developmental Disorders*, vol. 31, no. 1, pp. 5–17, Feb 2001. [Online]. Available: https://doi.org/10.1023/A:1005653411471

[19] M. H. Davis, “Measuring individual differences in empathy: Evidence for a multidimensional approach,” *J. Pers. Soc. Psychol.*, vol. 44, no. 1, pp. 113–126, 1983.

[20] V. Taksic, “The importance of emotional intelligence(competence) in positive psychology,” in *Proc. The First International Positive Psychology Summit*, 2002.

[21] P. T. Costa and R. R. McCrae, *Revised NEO Personality Inventory (NEO-PIR) and NEO Five Factor Inventory (NEO-FFI) professional manual*. Psychological Assessment Resources, 1992.

[22] H. Suematsu, S. Nomura, and M. Wada, *Handbook of TEG. 2nd edition [in Japanese]*. Tokyo: Kaneko-shobo, 1993.

[23] S. Watanabe, “Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory,” *J. Mach. Learn. Res.*, vol. 11, pp. 3571–3594, 2010.

[24] A. Vehtari, A. Gelman, and J. Gabry, “Practical bayesian model evaluation using leave-one-out cross-validation and waic,” *Statistics and Computing*, vol. 27, no. 5, pp. 1413–1432, 2017.

[25] H. Gunes and B. Schuller, “Categorical and dimensional affect analysis in continuous input: Current trends and future directions,” *Image Vision Comput.*, vol. 31, no. 2, pp. 120–136, 2013.

[26] T. Bänziger, D. Grandjean, and K. R. Scherer, “Emotion recognition from expressions in face, voice, and body: the multimodal emotion recognition test (MERT),” *Emotion*, vol. 9, pp. 691–704, 2009.

[27] T. K. Koo and M. Y. Li, “A guideline of selecting and reporting intraclass correlation coefficients for reliability research,” *Journal of Chiropractic Medicine*, vol. 15, no. 2, pp. 155 – 163, 2016.

[28] D. V. Cicchetti, “Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology,” *Psychological Assessment*, vol. 6, no. 4, pp. 284–290, 1994.

[29] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, “A multimodal database for affect recognition and implicit tagging,” *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 42–55, 2012.

[30] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, “Deap: A database for emotion analysis ;using physiological signals,” *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 18–31, 2012.

[31] H. P. Martinez, Y. Bengio, and G. N. Yannakakis, “Learning deep physiological models of affect,” *IEEE Comp. Intell. Magazine*, vol. 8, no. 2, pp. 20–33, 2013.